Review of Error Rate and Computation Time of Clustering Algorithms on Social Networking Sites

Jaskaranjit Kaur
M. Tech,
CSE
DAV University,
Jalandhar

Gurpreet Singh
Research Scholar
(Ph.D, Computer Sc. Engg.)
Pacific Academy of Higher Education and
Research University, Udaipur

ABSTRACT

Data mining is a method of finding useful patters from large volumes of data. It is an extension of traditional data analysis and statistical approaches. Data Clustering is a task of grouping a set of items or objects into subsets (called clusters). It is an algorithm to discover the similarity between objects in the same class (intraclass similarity) and minimizing the similarity between objects of different classes (interclass similarity). This paper discusses the standard KMeans clustering algorithm and Kohonen Self Organizing Map(SOM) clustering algorithm using the Tanagra datamining tool .These algorithms are applied on facebook dataset i.e which type of information is shared by university students on facebook.And that information is then used for product marketing purposes. And according to our analysis SOM gives best result with high accuracy and less computational time.

Keywords

Cluster Analysis, K-Means algorithm, Kohonen SOM algorithm, Tanagra Tool.

1. INTRODUCTION

Data mining is a multidisciplinary field which draws work from areas like database technology, statistics, machine learning , pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization. Data mining is a way of analyzing datasets in order to find out unsuspected relationships and to summarize the data in better ways that are both understandable and useful to the data owner. The summaries, relationships and groups that we get through a data mining process are referred to as patterns. It is a process of discovering hidden knowledge from large volumes of raw data. The knowledge must be new and appropriate so that one can use it. Data mining finds these patterns, groups and relationships using data analysis tools. We have two models in data mining:

- Predictive models, in which we use data with known results to build a model that can be used to explicitly predict values.
- Descriptive models, which describe patterns in existing data.

Clustering is an unsupervised learning technique which divides data items into a number of groups, such that items in the same cluster are more similar and items in different clusters are dissimilar, on the basis of some condition or criteria. It is different from supervised learning where the training examples are provided with class labels which describes the membership of every example to the appropriate class. In clustering we have no pror information about the classes.

Based on the information found in the data describing the objects or their relationships cluster analysis groups objects or events. Clustering is a tool to analyze the data. Its objective is to distribute the objects or events into groups, such that the members of the same cluster have strong degree of association than the members of different clusters. So in this way each cluster describes the class to which its members belong [5].

This paper includes two clustering algorithms: K-means and Kohonen SOM clustering algorithm.

2. CLUSTERING TECHNIQUES

2.1 K-Means Clustering

It is an algorithm to group objects based on attributes/features of the object into K number of groups. K is any positive integer. The clusters are formed on the basis of minimizing the sum of squares of distances between data and the corresponding cluster centroid.

K-Means is an unsupervised learning algorithm which clusters only numerical data in an iterative manner. The main basic concept behind the algorithm is to define k centroids, one for each cluster. As different locations of k causes different cluster result so centroids should be positioned in a schemind way. Then to find the k centroids, the K-Means clustering algorithm will iteratively cluster data and assign each object to the nearest centroid. The centroid here is the mean of the coordinates of the objects in the cluster. Then, the k centroids will change their positions step by step until no further changes occur [6]. The k-means algorithm is as follows [7]:

1. Select k points as initial centroids (randomly generated vectors can also be used).

2. Calculate the distance from each cluster centroid to each point.

3. Assign each point to the nearest cluster.

4. Calculate new cluster centroid, where each new centroid is the mean of all points in that cluster.

5. Repeat steps 2-4 until a stopping condition is reached.

2.2 Kohonen SOM Clustering

Self Organizing Map (SOMs) is an unsupervised learning. The basic notion of a SOM is to map the data patterns onto a n-dimensional grid of units. The grid act as an output space and the space where the data patterns are act as an input space.This mapping tries to preserve topological relations, i.e., patterns that are close in the input space will be mapped to units that are close in the output space and vice-versa [12]. It is a way to find a good mapping from high dimensional input space to the 2-D representative of the nodes. One way to use SOMs for clustering is to consider the objects in the input space represented by the same node grouped into cluster in the output space. At the training process, each object in the input is presented to the map and the best matching node is identified. After the locations of nodes are tuned, they will create meaningful coordinate system for the input features [6].

Algorithm for Kohonen's Self Organizing Map [1]

- Assume output nodes are connected in an array (usually 1 or 2 dimensional)
- Assume that the network is fully connected - all nodes in input layer are connected to all nodes in output layer.
- Use the competitive learning algorithm as follows:

1. Randomly choose an input vector x

2. Determine the "winning" output node i, where wi is the weight vector connecting the inputs to output node i.

w_i x >= w_k x only if the weights are normalized.

$$|\omega_i-X|\leq|\omega_K-X| \quad \forall k$$

Given the winning node i, the weight update is

$$\boldsymbol{\omega_k(new) = \omega_k(old) + \Delta\omega_k(n)}$$

where$\Delta\omega_k(n)$ represesnts the change in weight.

3. TANAGRA TOOL

TANAGRA is a free DATA MINING tool for data analysis and research purposes. This software is the advancement of SIPINA. SIPINA implements various supervised learning algorithms, mainly the visual construction of decision trees. TANAGRA is more powerful than it as it contains some supervised learning and also other paradigms such as clustering, parametric and nonparametric statistics, association rule, feature selection, factorial analysis and construction algorithms [4]. The main objective of Tanagra is to give researchers and students an easy data mining software, in compliance to the present work of the software development in this domain and allow to analyze data either real or synthetic. The second objective of TANAGRA is to allow researchers to easily add their datamining algorithms or methods and compare their performance with other methods. The third objective is to help developers in diffusing a probable methodology for building this type of software. Developers can take benefit to look how this kind of software is built, the main steps to follow in developing the project, the kind of problems which we have to avoid during development and which kind of tool and code libraries to use. So in this manner, Tanagra can be considered as an educational tool for learning programming techniques [4].

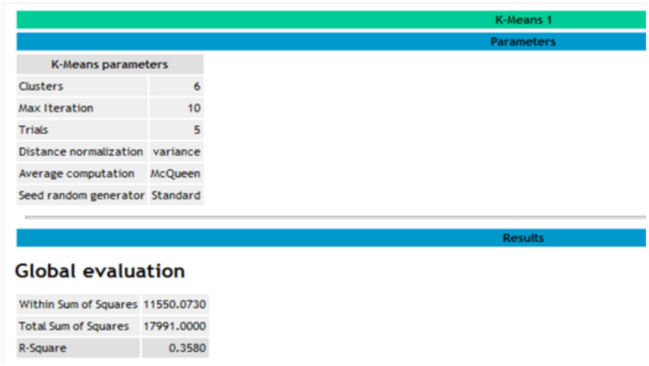
4. DATASET AND RESULTS

4.1 Data Set and Screenshots

The dataset used is a “Facebook” dataset. It has approximately 10 attributes and 1999 instances. The type of information shared among students is taken as a class attribute.

Table 1: Attributes of Data Set

Gender	Continuous
Age	Continuous
Area of Education	Continuous
Education	Continuous
Products	Continuous
No. of Facebook Friends	Continuous
No. of hours used	Continuous
No. of Groups joined	Continuous
No. of Social Networking sites joined	Continuous
Information Shared	Discrete



K-Means 1

Parameters

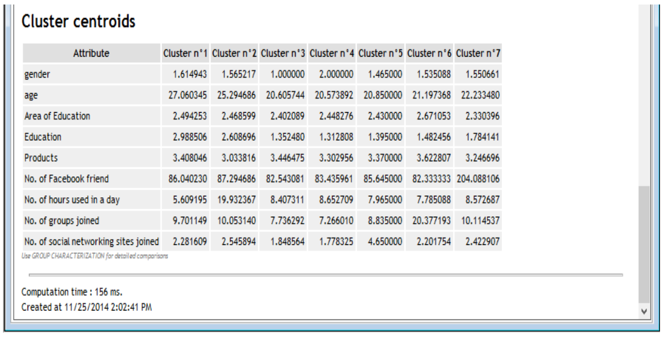
K-Means parameters	
Clusters	6
Max Iteration	10
Trials	5
Distance normalization	variance
Average computation	McQueen
Seed random generator	Standard

Results

Global evaluation

Within Sum of Squares	11550.0730
Total Sum of Squares	17991.0000
R-Square	0.3580

Figure 1. Implemented K-Means Clustering Algorithm with 6 clusters with error rate 0.3580



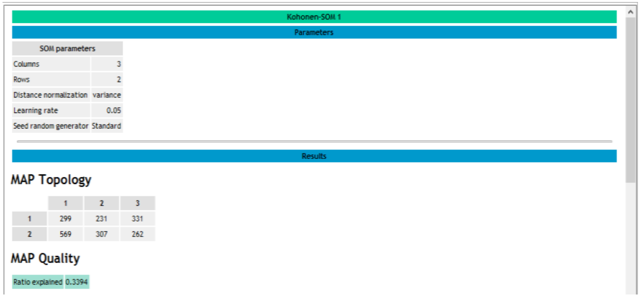
Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3	Cluster n°4	Cluster n°5	Cluster n°6	Cluster n°7
gender	1.614943	1.565217	1.000000	2.000000	1.465000	1.535088	1.550661
age	27.060345	25.294686	20.605744	20.573892	20.850000	21.197368	22.233480
Area of Education	2.494253	2.468599	2.402089	2.448276	2.430000	2.671053	2.330396
Education	2.988506	2.608696	1.352480	1.312808	1.395000	1.482456	1.784141
Products	3.408046	3.033816	3.446475	3.302956	3.370000	3.622807	3.246696
No. of Facebook friend	86.040230	87.294686	82.543081	83.435961	85.645000	82.333333	204.088106
No. of hours used in a day	5.609195	19.932367	8.407311	8.652709	7.965000	7.785088	8.572687
No. of groups joined	9.701149	10.053140	7.736292	7.266010	8.835000	20.377193	10.114537
No. of social networking sites joined	2.281609	2.545894	1.848564	1.778325	4.650000	2.201754	2.422907

Use GROUP CHARACTERIZATION for detailed comparisons

Computation time : 156 ms.
Created at 11/25/2014 2:02:41 PM

Figure 2. Implemented K-Means Clustering Algorithm with 6 clusters with computation time 156ms



Kohonen-SOM 1

Parameters

SOM parameters	
Columns	3
Rows	2
Distance normalization	variance
Learning rate	0.05
Seed random generator	Standard

Results

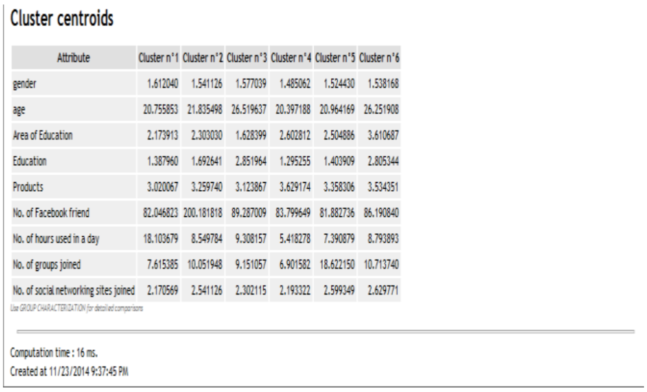
MAP Topology

	1	2	3
1	299	231	331
2	569	307	262

MAP Quality

Ratio explained	0.3394

Figure 3. Implemented Kohonen SOM Clustering Algorithm and generate MAP Topology with 6 clusters, with Error rate 0.3395



Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3	Cluster n°4	Cluster n°5	Cluster n°6
gender	1.612040	1.541126	1.577039	1.485062	1.524430	1.538168
age	20.755853	21.835498	26.519637	20.397188	20.964169	26.251908
Area of Education	2.173913	2.303030	1.628399	2.602812	2.504886	3.610687
Education	1.387960	1.692641	2.851964	1.295255	1.403909	2.805344
Products	3.020067	3.259740	3.123867	3.629174	3.358306	3.534351
No. of Facebook friend	82.046823	200.181818	89.287009	83.799649	81.882736	86.190840
No. of hours used in a day	18.103679	8.549784	9.308157	5.418278	7.390879	8.793893
No. of groups joined	7.615385	10.051948	9.151057	6.901582	18.622150	10.713740
No. of social networking sites joined	2.170569	2.541126	2.302115	2.193322	2.599349	2.629771

Use GROUP CHARACTERIZATION for detailed comparisons

Computation time : 16 ms.

Created at 11/23/2014 9:37:45 PM

Figure 4. Implemented Kohonen SOM Clustering Algorithm with 6 clusters and gives Computation time 16ms

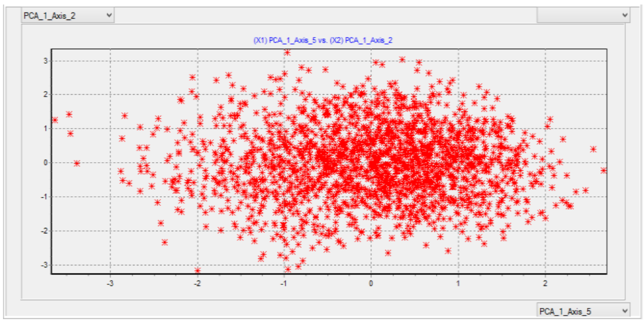


Figure 5. Data before applying Clustering Algorithms

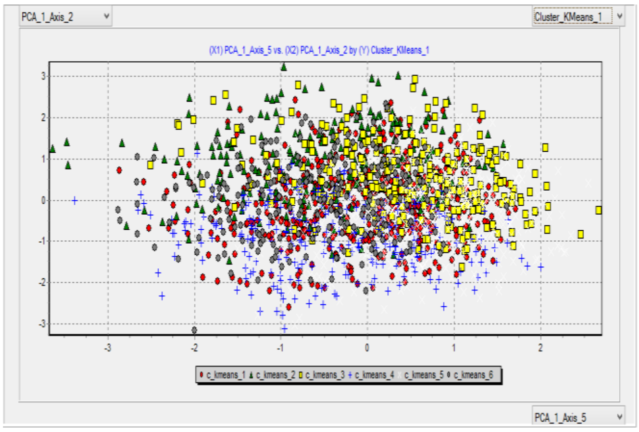


Figure 6. Six types of clusters after applying k-means clustering algorithm

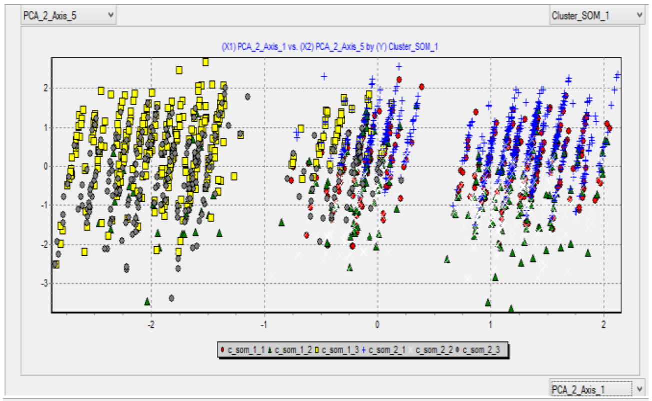


Figure 7. Six types of clusters after applying Kohonen SOM clustering algorithm

4.2 Result

After implementation of these algorithms on Facebook dataset, the following results are obtained:

Table 2: Results of both algorithms

Parameters	K-Means	Kohonen SOM
No. of clusters	6	6
Error Rate	0.3580	0.3395
ComputationTime	156ms	16ms

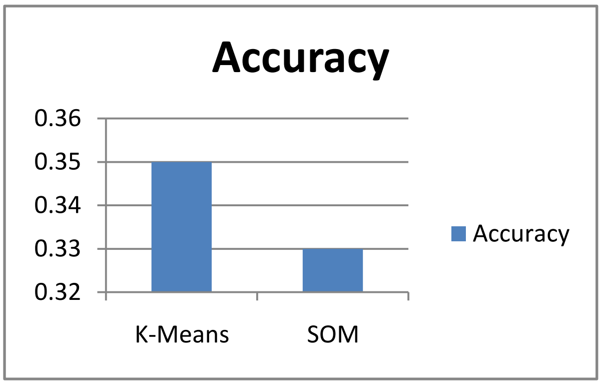


Figure 8. Graphical representation of Accuracy (K-Means=0.3580 and SOM=0.3395)

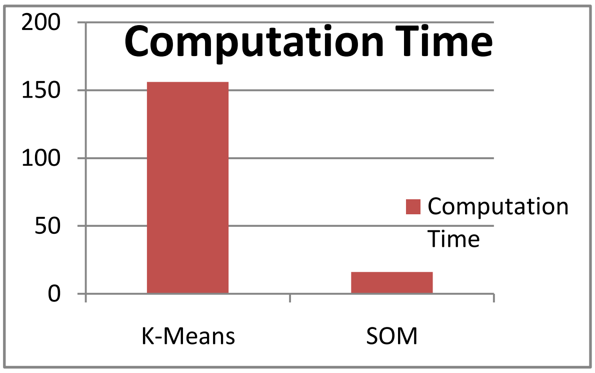


Figure 9. Graphical representation of Computation Time(K-Means=156ms and SOM=16ms)

5. CONCLUSION

The data based upon the type of information shared is collected and validated. Efficient clustering algorithms (K-Means and Kohonen SOM) are applied to finalize the number of clusters which resulted in six qualified clusters. Out of these SOM gives more accuracy. If marketers are interested in expanding the market, they should target to promote the products which are of interests to students by measuring their type of information shared. This clustering information will provide direct benefits to marketers, especially when students have large number of friends in the network where outgrowths for commercial can possibly be gained. The marketers get the opportunity to promote their products as these specific groups have high purchasing power while maintaining close contact with their correspondents.

6. REFERENCES

[1] Shaina Dhingra , RimpleGilhotra,Ravishanker, "Comparative Analysis of Kohonen-SOM and K-Means data mining algorithms based on Academic Activities", 2013 International Journal of Computer Applications(0987-8887)

[2] RichaDhiman, ShevetaVashisht, "A Cluster analysis and Decision Tree Hybrid Approach in Data Mining to Describe Tax Audit", International Journal of Computers & Technology Volume 4 No. 1, Jan-Feb, 2013

[3] Saurabh Shah,Manmohan Singh, "Comparison of A Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid algorithm", 2012 IEEE International Conference on Communication Systems and Network Technologies.

[4] Y. Ramamohan, K. Vasantharao , C. KalyanaChakravarti , A.S.K.Ratnam, "A Study of Data Mining Tools in Knowledge Discovery Process", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012

[5] Shi Na , Liu Xumin, Guan yong , "Research on k-means Clustering Algorithm An Improved k-means Clustering Algorithm", 2010 IEEE Third International Symposium on Intelligent Information Technology and Security Informatics.

[6] SuwimonVongsingthong,NawapornWisitpongphan,"Classification of University Students' Behaviors in Sharing Information on Facebook", 2014 IEEE 11th International Joint Conference on Computer Science and Software Engineering (JCSSE)

[7] Shalove Agarwal, Shashank Yadav, Kanchan Singh, "K-means versus K-means ++ Clustering Technique", 2012 IEEE Second International Workshop on Education Technology and Computer Science

[8] W.-L. C. T.-H. Lin, "A Cluster-Based Approach for Automatic Social Network Construction", 2010 IEEE Second International Conference on Social Computing (SocialCom), pp. 601 - 606 2010.

[9] Wei-Lun Chang, Tzu-Hsiang Lin, "A Cluster-Based Approach for Automatic Social Network Construction", 2010 IEEE International Conference on Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust

[10] S. N. Alsaleh, R. , Yue Xu, "Grouping people in social networks using a weighted multi-constraints clustering method ", 2012 IEEE InternationalConference on Fuzzy Systems (FUZZ-IEEE), pp. 1-8, 2012

[11] Ying He, Tian-Jin Feng, Jun-Kuo Cao, Xiang-Qian Ding Y, Ing-Hui Zhou, "Research on Some Problems in the Kohonen SOM Algorithm" , IEEE Proceedings of the First International Conference on Maclune Learning and Cybernetics, Beijing, 4-5 November 2002.

[12] Fernando Bacao, Victor Lobo, Marco Painho, "Self-organizing Maps as Substitutes for K-Means Clustering", Springer-Verlag Berlin Heidelberg 2005.